

# Model-free reinforcement learning operates over information stored in working-memory to drive human choices

Carolina Feher da Silva<sup>\*1</sup>, Yuan-Wei Yao<sup>2</sup>, and Todd A. Hare<sup>3,4</sup>

<sup>1</sup>Department of General Physics, Institute of Physics, University of São Paulo

<sup>2</sup>State Key Laboratory of Cognitive Neuroscience and Learning and IDG/McGovern Institute for Brain Research, Beijing Normal University

<sup>3</sup>Laboratory for Social and Neural Systems Research, Department of Economics, University of Zurich

<sup>4</sup>Zurich Center for Neuroscience, University of Zurich and ETH

February 10, 2017

## Abstract

Model-free learning creates stimulus-response associations, but are there limits to the types of stimuli it can operate over? Most experiments on reward-learning have used discrete sensory stimuli, but there is no algorithmic reason to restrict model-free learning to external stimuli, and theories suggest that model-free processes may operate over highly abstract concepts and goals. Our study aimed to determine whether model-free learning can operate over environmental states defined by information held in working memory. We compared the data from human participants in two conditions that presented learning cues either simultaneously or as a temporal sequence that required working memory. There was a significant influence of model-free learning in the working memory condition. Moreover, both groups showed greater model-free effects than simulated model-based agents. Thus, we show that model-free learning processes operate not just in parallel, but also in cooperation with canonical executive functions such as working memory to support behavior.

---

\*Corresponding author

# 1 Introduction

Reinforcement learning theory and the computational algorithms associated with it have been extremely influential in the behavioral, biological, and computer sciences. Reinforcement learning theory describes how an agent learns by interacting with its environment [1]. In a typical reinforcement learning paradigm, the agent selects an action and the environment responds by presenting rewards and taking the agent to the next situation, or state. A reinforcement learning algorithm determines how the agent changes its action selection strategy as a result of experience, with the goal of maximizing future rewards. Depending on how algorithms accomplish this goal, they are classified as model-free or model-based [1]. Model-based algorithms acquire beliefs about how the environment generates outcomes in response to their actions and select actions according to their predicted consequences. By contrast, model-free algorithms generate a propensity to perform, in each state of the world, actions that were more rewarding in previous visits to that environmental state. Model-free reinforcement learning algorithms are of considerable interest to behavioral and biological scientists, in part because they offer a compelling account of the phasic activity of dopamine neurons, but also more generally can explain many observed patterns of behavior in human and non-human animals [2, 3, 4, 5, 6, 7].

A key concept in reinforcement learning theory is the environmental state. Typically, empirical tests of reinforcement learning algorithms use discrete sensory stimuli to define environmental states. However, there is no theoretical or algorithmic constraint to define the states of the environment exclusively by sensory stimuli. State definitions may also include the agent's internal stimuli, such as its memory of past events, thirst or hunger level, or even subjective characteristics such as happiness or sadness [1]. Thus, model-free reinforcement learning might operate over a wide variety of both external and internal factors.

Indeed, recent work suggests that model-free learning algorithms can support a large set of cognitive processes and behaviors beyond the formation of habitual response associations with discrete sensory stimuli [8, 9, 10]. For instance, it has been proposed that the model-free system can perform the action of selecting a goal for goal-directed planning [11] or conversely that a model-based decision can trigger a habitual action sequence [12, 13, 14, 15]. Model-free algorithms have also been suggested to gate working memory [16]. However, many of these important theoretical proposals about model-free algorithms have not been directly tested empirically.

Here, we determine the ability of model-free reinforcement learning algorithms to operate over states defined by information held in working memory, an internal state. Specifically, we use an experimental paradigm and computational modeling framework designed to dissociate model-

free from model-based influences on behavior [17] to test if temporally separated sequences of individually uninformative cues can drive model-free learning and behavior. If an agent can store the elements of a temporal sequence in its memory to form a unique and predictive cue and use the memorized information as the state definition, then, theoretically, it can use model-free algorithms to learn the associations between a specific sequence of *individually uninformative cues* and action outcomes [18].

Our approach has several important facets. First, we use an experimental paradigm that allows us to determine not only if our participants learn from information in working memory, but also whether that learning is supported by model-based or model-free algorithms. Second, the cues in our temporal sequences are individually uninformative; in other words, any single cue in isolation provides no information about which response is correct. It is well-known that model-free algorithms can shift response associations to the earliest occurring predictor of the correct response in a temporal sequence of informative cues and can integrate predictive information across individual cues. Neither of these mechanisms is possible in our paradigm because the individual cues themselves contain no information about the previous or subsequent cues or which response is best.

Temporal pattern learning is a fundamental and early developing human cognitive ability. It allows people to form predictions about what will happen from what has happened and select their actions accordingly. Humans can learn patterns both explicitly and implicitly in the absence of specific instructions or conscious awareness [19]. Moreover, they can do so as early as two months of age [20]. In fact, people identify patterns even when, in reality, no pattern exists [21]. These empirical results together with the theoretical potential for model-free learning to operate over internal stimuli suggest that temporal pattern learning could be supported by model-free processes. However, to date, studies of reinforcement learning and decision making have focused primarily on tasks in which the relevant stimuli are presented simultaneously just prior to or at the time of decision-making, or on implicit motor sequence learning, wherein participants learn a sequence of movements automatically, without full awareness (for instance, 22, 23, 24, 25, 26). Thus, the degree to which model-free processes do in fact operate over temporal sequences or any other information stored in working memory has not yet been directly tested and compared with model-free learning from traditionally employed external, static environmental cues.

Here, we directly test whether model-free processes can access and learn from information stored in working memory. We adapted a decision-making paradigm originally developed by Daw et al. [17] that can behaviorally dissociate the influence of model-free and model-based learning on choice. The task was performed by two groups of human participants either in a simultaneous

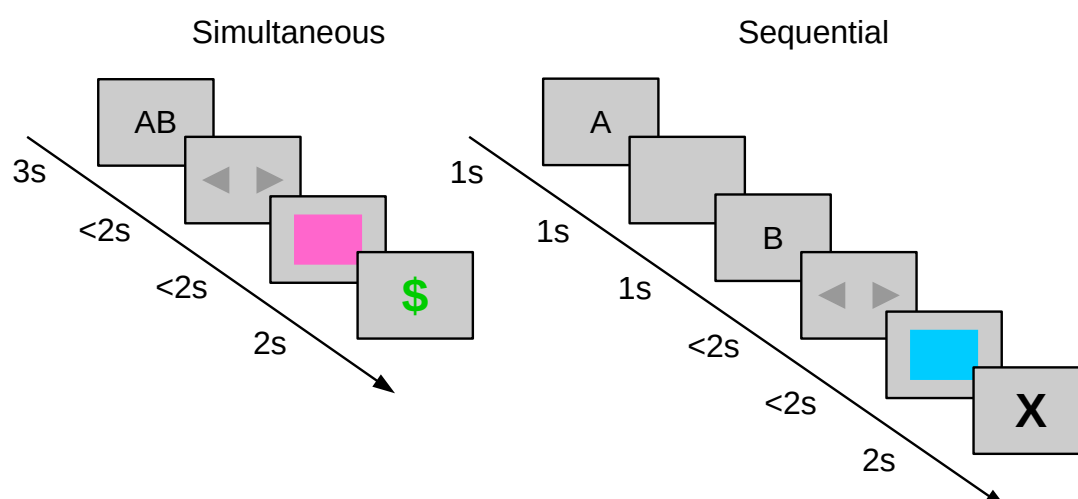


Figure 1: Timelines of events in a trial. The two symbols that represent the initial state are presented simultaneously in the simultaneous condition (left) and separately as a temporal sequence in the sequential condition (right). In this example, AB is the initial state. The simultaneous condition participant goes to the pink final state and receives a reward (signaled by the green \$ symbol). The sequential condition participant goes to the blue final state and does not receive a reward (signaled by the black X symbol).

condition (i.e. static and external), wherein visual stimuli were presented simultaneously, or in a sequential condition, wherein the same visual stimuli were presented as a temporal sequence that required working memory processing. We also simulated a series of experiments in which artificial model-based agents whose behavioral processes we determined were compared to the human participants. Our analysis indicates that our temporal sequences, and consequently information stored in working memory, can trigger model-free learning. Moreover, we found no evidence that the degree to which model-free learning influenced behavior differed between conditions in which environmental states were defined by external sensory stimuli compared to those defined by internal representations stored in working memory. Our findings support the theoretical proposition that model-free learning can act on stimuli internally represented in working memory as well as on external ones.

## 2 Results

### 2.1 Determining model-free and model-based influences on choice behavior

Forty-one young human participants completed a behavioral task adapted from Daw et al. [17]. In our task, participants began each trial in a randomly selected initial state represented by one

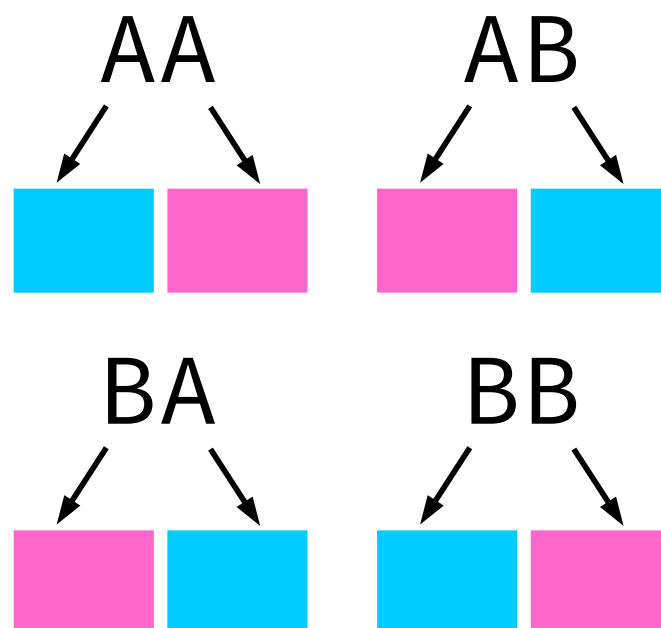


Figure 2: Common state transitions in the behavioral task's model. These graphics highlight the uninformative nature of each single element (i.e. A or B symbols) in the simultaneous or sequential cues. Knowledge of only the first or final element of the combined cue provides no indication of how likely the right and left responses are to lead to a specific state.

of four possible sequences of two symbols: AA, AB, BA, or BB (Figure 1). At this initial state, participants chose one of two possible actions: going left or going right. They were then taken to one of two possible final states, the blue state or the pink state. If they had gone left, they were taken with 0.8 probability to the final state given by the rule AA → blue, AB → pink, BA → pink, BB → blue or with 0.2 probability to the other final state. If they had gone right, they were taken with 0.8 probability to the final state *not* given by the previous rule or with 0.2 probability to the other final state. The common (most probable) transitions between the initial and final states are shown in Figure 2. To predict the final state accurately, participants had to know both elements of the sequence. If they knew only one, the final state might have been either blue or pink with 0.5 probability and they would not be able to perform above chance. This feature is key and separates our work from others in which each element of a sequence is predictive on its own.

One of the final states delivered a monetary reward with 0.7 probability and the other with 0.3 probability. The optimal strategy was to always select the action that led with 0.8 probability to the final state with 0.7 reward probability. Initially, participants were instructed to learn the common transitions between the initial and final states in the absence of rewards. They were told that each final state might be rewarded with different probabilities, but not what the probabilities were nor that they were fixed. The task comprised 250 trials and participants received the total reward they obtained at the end.

115 Twenty-one participants were randomly allocated to a simultaneous condition and twenty to a  
116 sequential condition (Figure 1). In the simultaneous condition, both symbols that represented the  
117 initial state were displayed simultaneously on the screen. In the sequential condition, each symbol  
118 was displayed consecutively by itself, as a temporal sequence. The specific objective of this study  
119 was to determine if participants in the sequential condition could use states represented in working  
120 memory to learn the task in a model-free way or if their learning was necessarily model-based. The  
121 simultaneous condition is already known to support model-free learning as well as model-based  
122 learning [17, 27, 28, 29, 30]. We thus sought to determine the difference between the standard  
123 simultaneous and working-memory dependent sequential conditions.

124 The two-stage task we used can differentiate between model-free and model-based learning  
125 because algorithms that implement them make different predictions about how a reward received  
126 in a trial impacts a participant's choices in subsequent trials. The SARSA ( $\lambda = 1$ ) model-free  
127 algorithm learns this task by strengthening or weakening associations between initial states and  
128 initial-state actions depending on whether the action is followed by a reward or not [1]. Therefore,  
129 it simply predicts that an initial-state action that resulted in a reward is more likely to be repeated  
130 in the next trial with the same initial state [17]. On the other hand, the model-based algorithm  
131 considered in this study uses an internal model of the task's structure to determine the initial-  
132 state choice that will most likely result in a reward [17]. To this end, it considers which final state,  
133 pink or blue, was most frequently rewarded in recent trials and selects the initial-state action, left  
134 or right, that will most likely lead there. Therefore, the model-free algorithm predicts that the  
135 participant will choose the mostly frequently rewarded *action* in past trials with the same initial  
136 state, while the model-based algorithm predicts that the participant will choose the action with the  
137 highest probability of leading to the mostly frequently rewarded *final state* in past trials, regardless  
138 of their initial states.

139 The model-free and model-based algorithms thus generate different predictions about the *stay*  
140 *probability*, which is the probability that in the next trial with the same initial state the participant  
141 will stay with their previous choice and take the same initial-state action. For instance, if in a  
142 given trial whose initial state was AA the participant chose left, and in the next trial with AA  
143 as the initial state the participant also chose left, this was considered a stay. The model-free  
144 prediction is that the stay probability will increase if the previous trial with the same initial state  
145 was rewarded and decrease if it was not. The model-based prediction, on the other hand, depends  
146 on the transition structure of the task and how the estimated reward probabilities of the two final  
147 states have changed since the previous trial with the same initial state (see Methods for a detailed  
148 description of how model-based predictions were calculated).

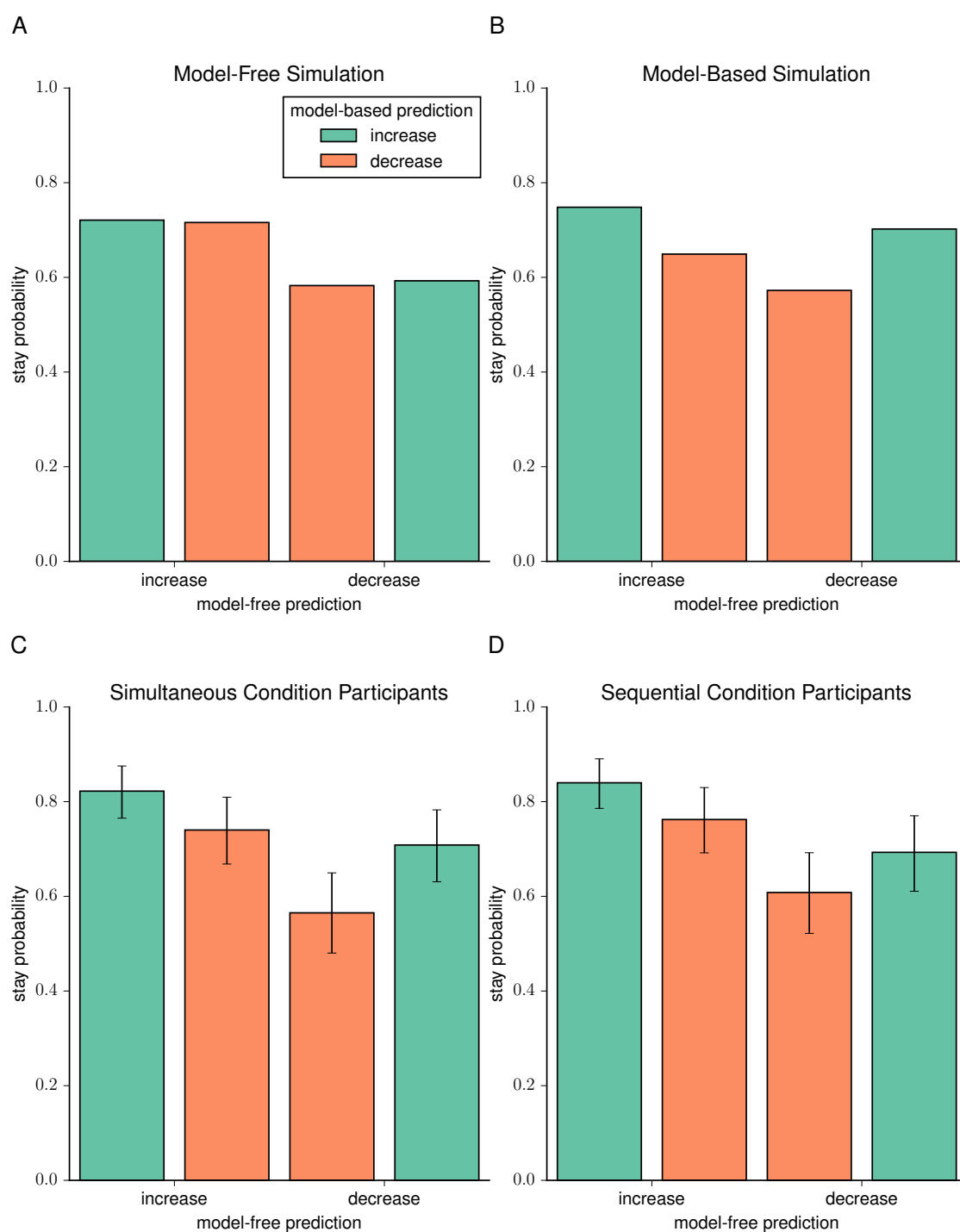


Figure 3: Stay probabilities for simulated agents and human participants as a function of model-free and model-based predictions. The bar graphs show the choice probabilities derived from the logistic regressions as a function of model-free (separated along the  $x$ -axis) and model-based predictions (indicated by the color of the bars). The four panels demonstrate the behavior of **A**) model-free simulations ( $N = 10,000$ ), **B**) model-based simulations ( $N = 10,000$ ), **C**) human participants in the simultaneous condition ( $N = 21$ ) and **D**) human participants in the sequential condition ( $N = 20$ ). Error bars on the data from human participants represent the 95% highest density interval.

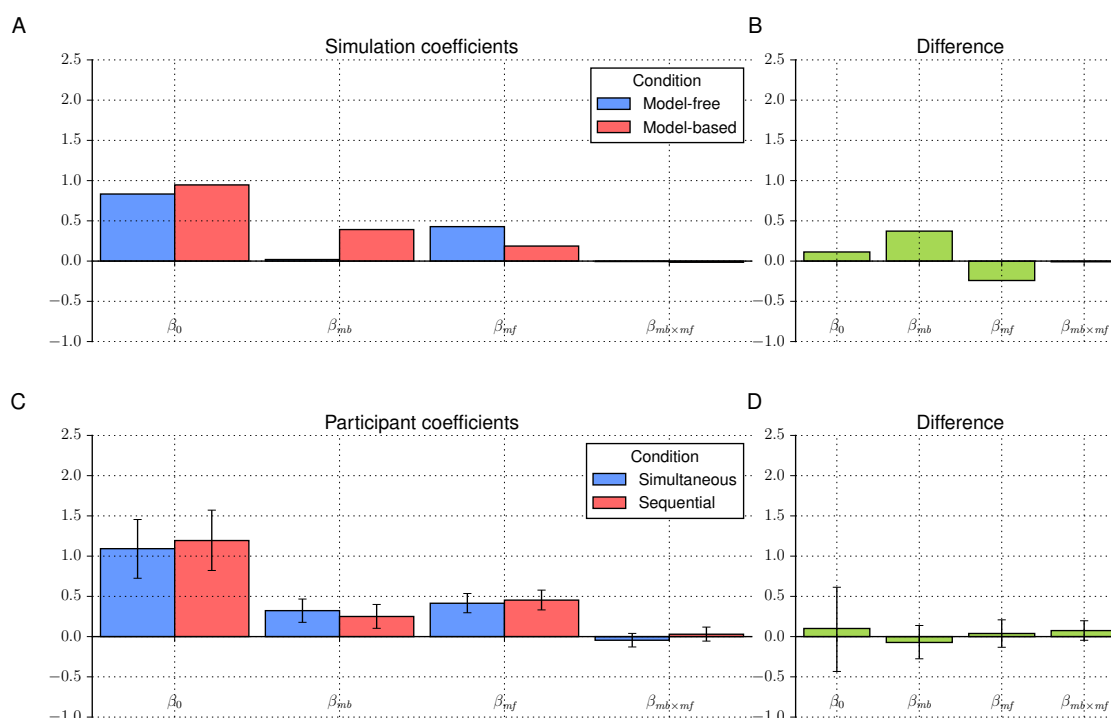


Figure 4: The relative effects of model-free and model-based learning on choice behavior. **A)** The regression coefficients from a logistic regression on stay vs switch choices for the model-free (blue;  $N = 10,000$ ) and the model-based simulations (purple;  $N = 10,000$ ). **B)** The difference between model-based and model-free simulation coefficients (i.e. red minus blue from panel A). **C)** Logistic regression coefficients from the same model used for panel A, but here estimated on choices from the simultaneous (blue;  $N = 21$ ) and sequential condition (purple;  $N = 20$ ) participants. **D)** The difference between sequential and simultaneous condition participants' coefficients (i.e. red minus blue from panel C). In all panels,  $\beta_0$  is the logistic regression's intercept,  $\beta_{mb}$  is the model-based coefficient,  $\beta_{mf}$  is the model-free coefficient, and  $\beta_{mb \times mf}$  is the coefficient of the interaction between the model-based and model-free effects. Error bars on the data from human participants represent the 95% highest density interval.

We simulated model-free and model-based agents performing this task for comparison with the behavior of human participants in each condition. In all cases, we analyzed the data using Bayesian hierarchical logistic regression analyses. The correspondence between theoretical predictions of the model-free and model-based algorithms and choices of the simulated agents are shown in the top row of Figure 3. The correspondence between theoretical predictions of the model-free and model-based algorithms and choices of the human participants in each experimental condition are shown in the bottom row of Figure 3.

In addition to examining the stay choice probabilities, we directly tested the degree to which the human participants' and simulated agents' choices were influenced by model-based and model-free signals. The coefficients of these logistic regression analyses are shown in Figure 4. The positive value of the intercept  $\beta_0$  indicates that the stay probabilities tended to be above 0.5, i.e., simulated agents and human participants were more likely to repeat their previous choice than switch to the



other choice in the next trial with the same initial state. This is also visible in the stay probabilities shown in Figure 3. The coefficients for the regressions on the simulated agents' choices show the expected pattern with the model-free and model-based coefficients primarily determining behavior for the model-free and model-based agents, respectively (Figure 4A) and differing substantially between agent types (Figure 4C).

In the human participants, behavior was influenced by both model-based and model-free processes regardless of whether the states were defined by external sensory cues or internal working-memory representations. The model-based and model-free coefficients,  $\beta_{mb}$  and  $\beta_{mf}$ , were positive for both the simultaneous and sequential conditions with 0.999 posterior probability (Figure 4B). The model-based coefficient was 0.32 (95% highest density interval [0.18, 0.47]) for the simultaneous condition and 0.25 (95% HDI [0.10, 0.40]) for the sequential condition, and the model-free coefficient was 0.41 (95% HDI [0.30, 0.54]) for the simultaneous condition and 0.45 (95% HDI [0.33, 0.58]) for the sequential condition. The differences between the sequential and simultaneous conditions were  $-0.07$  (95% HDI  $[-0.28, 0.14]$ ) for the model-based coefficient and  $0.04$  (95% HDI  $[-0.13, 0.21]$ ) for the model-free coefficient. The posterior probability that the model-free coefficient is smaller in the sequential group than in the simultaneous group is 0.32, and the posterior probability that the model-based coefficient is greater in the sequential group than in the simultaneous group is 0.24 (Figure 4D). Thus, we find no evidence that sequentially presented, working-memory-dependent state cues shift the balance of model-based and model-free effects on choice behavior compared to traditional, static, external cues.

The model-based predictions in our two-stage decision task differ from those reported in previous work using similar tasks. In the version of the two-stage task used by Daw et al. [17], the model-based prediction is that the heights of the two orange bars should be (nearly) equal to one another and that the heights of the two green bars should be (nearly) equal as well (note, that the precise prediction depends on the exact parameterization of the model). However, in our task the model-based prediction includes a reward effect. Consequently, we find that the stay probabilities in the model-based agent simulations are influenced by the outcome of the most recent trial with the same initial state as can be seen in the differences in magnitude between the two inner, orange bars as well as the two outer, green bars in Figure 3B. Specifically, if the previous matching-state trial was rewarded, then the stay probabilities are greater than if it was not rewarded. This reward effect in the model-based choices is similar to a model-free effect, but not identical because the model-based value updating procedure incorporates the transition probabilities while the model-free algorithm does not. However, the reward-effect does lead to model-free-like choice patterns in the data that result in a small, but significant model-free coefficient in the logistic regressions

on model-based agents' choices (Figure 4A). Therefore, small model-free-like patterns in the stay probabilities do not necessarily indicate the influence of model-free learning, because we know the model-based agents do not use this learning algorithm. We directly address the potential for spurious effects mimicking the influence of a model-free algorithm in our human participants in the working-memory-dependent sequential condition in the following paragraphs.

## 2.2 Direct comparisons between human participants and simulated model-based agents

Our goal was to test the hypothesis that working-memory-dependent, temporal patterns can be learned through a model-free process in humans. Given that our model-based simulations showed a reward effect that shared some properties with a model-free learning process, we sought to determine if the same results obtained by the human participants in the sequential condition, including the estimated model-free effect, could have been generated through the use of a model-based algorithm alone. Despite the fact that we find no evidence for differences in behavior between participants in the sequential and simultaneous conditions, this additional test is important because as Figures 3 and 4 show, the model-based simulated agents exhibited behavior that mimicked a model-free effect even though they operated solely on the basis of a model-based algorithm by design.

This raises the question, could purely model-based agents exhibit a model-free effect as large as the participants in the sequential condition? To this end, we fitted the model-based algorithm to the sequential condition results using a Bayesian hierarchical model. We then simulated 10,000 experiments in which we first created behavior for 20 simulated model-based agents (replacing the 20 sequential condition human participants) and then combined those data with that from the 21 human participants in the simultaneous condition and estimated the same hierarchical logistic regression on stay/switch choices described above and summarized in Figure 4. These 10,000 regressions give us a measure of what the coefficients in the sequential condition participants would be if they were purely model-based.

We found that while the simulated purely-model-based (PMB) agents showed a level of model-based influence comparable to participants in the sequential condition, the degree of model-free influence in PMB agents was substantially lower. The mean value of the model-based coefficient,  $\beta_{mb}$ , was 0.23 (95% HDI [0.05, 0.49]), which, as expected, is very close to the mean value of 0.25 from the sequential participants' behavior. Likewise, the mean difference across simulations between the PMB agents and the participants in the simultaneous condition for  $\beta_{mb}$  was -0.09

(95% HDI  $[-0.28, 0.15]$ ), similar to the  $-0.07$  value for the difference between the sequential and simultaneous conditions in humans. In contrast, the mean value of the model-free coefficient,  $\beta_{mf}$ , in the simulated agents was  $0.15$  (95% HDI  $[0.07, 0.23]$ ), and it was smaller than  $0.45$ , the mean value obtained for the sequential condition, in more than 99.9% (in fact all 10,000) of the simulated experiments. Furthermore, the mean difference between the PMB agents and the participants in the simultaneous condition for  $\beta_{mf}$  (i.e. the difference corresponding to Figure 4D) was  $-0.26$  (95% HDI  $[-0.34, -0.18]$ ), and more than 99.9% (in fact all 10,000) of simulated experiments yielded a difference for  $\beta_{mf}$  smaller than  $0.04$ , the observed difference between human participants in the sequential and simultaneous conditions. In summary, the model-free coefficient observed in the sequential condition is three times the size one would expect to see from a purely model-based agent, which strongly suggests that the observed effect is due to a true model-free influence and not mimicked by the reward-effect.

### 3 Discussion

In this study, we empirically tested the hypothesis that human participants can develop model-free associations between temporal sequences of stimuli stored in working memory and a motor response. To that end, we developed a behavioral task based on a previous decision-making paradigm that can determine the model-free and model-based influences on choice [17]. The participants in the simultaneous condition performed this task with the two visual symbols presented together simultaneously and those in the sequential condition performed it with the same two visual symbols presented as a temporal sequence that had to be held in working memory. The model-free effect estimated for the sequential condition was similar to the one estimated for the simultaneous condition and higher than that predicted by a purely model-based algorithm. Our results suggest that both model-based and model-free learning influenced the participants' choices whether they saw the entire set of stimuli at once or saw each stimulus by itself at separate times. Our study thus provides experimental support to proposed model-free algorithms of temporal pattern learning [18] and the view that model-free learning and habituation can be triggered by external or internal stimuli [8, 9, 10]

A key element of our experimental paradigm is that the individual symbols within each temporal sequence convey no information about the best response in isolation. This fact rules out the possibility that the sequential condition's model-free effect is due to an association between a single symbol in the sequence and a response rather than one between the entire sequence and a response. Each sequence element is completely uninformative by itself: it cannot predict reward delivery

above chance. Therefore, the task cannot be learned by simple stimulus-response associations with individual symbols in the temporal sequence.

Model-free learning processes support habit formation, and thus our results suggest that stimuli stored in working memory can trigger habitual responses. To the best of our knowledge, no study has yet tested for habituation to temporal sequences directly, using procedures such as contingency degradation or outcome devaluation. Although two-stage choice tasks similar to the one we use here have been reported to share construct validity with outcome devaluation measures of habitual responding [31], direct tests of outcome devaluation and contingency degradation following temporal sequence learning are still needed. If such additional tests show positive evidence for habituation, this would indicate that habits can be triggered by internally generated stimuli as well as by external ones. Conversely, if no evidence is found for habituation to temporal sequences, this would indicate that model-free learning processes can use internal stimuli, but do not necessarily produce habits. Experimental evidence already suggests that habits are not exclusively learned in a model-free way [32]; it may also be true that habituation involves additional mechanisms beyond the model-free caching of state-action-reward contingencies. Our study also raises the question of which neural systems are commonly versus distinctly recruited in order to learn from stimuli represented in working memory (e.g. temporal sequences) compared to purely external stimuli in a reinforcement learning task. While numerous studies have investigated the neural systems mediating reinforcement learning over externally presented stimuli (see 33 for a review), to date, only a single study has investigated brain activity involved in temporal pattern learning using fMRI [21]. However, the sequence of events in that study was random, and any pattern that occurred was spurious. Moreover, participants were required to respond to the stimuli instead of predicting them, and might thus be implicitly learning a motor sequence. It remains to be determined what brain regions support explicit learning from temporal sequences, or other stimuli held in working memory, and to what degree these systems overlap with those shown to underlie learning from external environmental cues. In conclusion, we have presented experimental evidence that temporal pattern learning, and consequently learning from internal stimuli held in working memory, can be model-free.

Our study has helped delineate the contexts that support model-free learning—a subject of current debate. Temporal pattern learning is a fundamental aspect of human cognition and model-free learning and habit formation are subjects of immediate relevance for research on typical learning as well as for the study of neuropsychiatric disorders ranging from addiction, obsessive-compulsive disorder, and Tourette syndrome to anxiety disorders and major depression [34]. It is thus important to continue investigating temporal pattern learning, including whether the model-free learning

of temporal sequences produces outcome-insensitive, habitual responses and how such learning is implemented in the brain.

## 4 Methods

### 4.1 Participants

Forty-one healthy young adults participated in the experiment, 21 (13 female) randomly assigned by a random number generator to the simultaneous condition and 20 (13 female) to the sequential condition. The inclusion criterion was speaking English and no participants were excluded from the analysis. The sample size was chosen by the precision for research planning method [35, 36], by comparing the estimated differences between participant groups in the logistic regression analysis with those between model-free and model-based simulated agents.

The experiment was conducted in accordance with the Zurich Cantonal Ethics Commission's norms for conducting research with human participants, and all participants gave written informed consent.

### 4.2 Task

The task's state transition model defines four possible initial states, which were randomly selected with uniform distribution in each trial and represented by four different stimuli, each composed of two symbols: AA, AB, BA, or BB. At the initial state, two actions were available to the participant: pressing the left or the right arrow keys. By pressing one of the keys, the participant was taken to a final state, which might be either the blue state or the pink state. If the left arrow key was pressed, the participant was taken to the final state given by the rule  $AA \rightarrow \text{blue}$ ,  $AB \rightarrow \text{pink}$ ,  $BA \rightarrow \text{pink}$ ,  $BB \rightarrow \text{blue}$  with 0.8 probability or to the other state with 0.2 probability; if the right arrow key was pressed, the participant was taken to the final state not given by the previous rule with 0.8 probability or to the other state with 0.2 probability. There was no choice of action at the final state, but participants were required to make a button press to potentially earn the reward. Each final state was rewarded according to an associated probability, which was 0.7 for one state and 0.3 for the other. The highest reward probability was associated with the blue state for half of the participants and to the pink state for the other half. Participants were told that each final state might be rewarded with different probabilities, but not what the probabilities were nor that they were fixed.

In contrast with our task design, in which the final states' reward probabilities were fixed, in

the original task design proposed by Daw et al. [17] the reward probabilities slowly drifted over time, because those authors were interested in the trade-off between model-based and model-free mechanisms, which is assumed to happen on the basis of their relative uncertainties. In this study we were interested instead in testing if model-free learning of temporal patterns is possible and keeping the task environment stable helps making the model-free associations stronger and more likely to influence choice [37, 38].

Participants were initially instructed to learn the common transitions between the initial and the final states in the absence of reward. Participants then performed the task defined by the model above in the simultaneous or sequential condition. Half of the participants were randomly allocated to the simultaneous condition and the other half to the sequential condition (Figure 1). In the simultaneous condition, both symbols that define the initial state were displayed simultaneously on the screen for 3 seconds. In the sequential condition, each symbol is an element of a sequence and each element was presented for 1 second, but never conjointly, and with a 1-second delay (blank screen) in between. Two triangles pointing left and right then appeared and the participant was given 2 seconds to make a decision about whether to press the left or the right arrow keys; if they did not press any keys, the word SLOW was displayed for 1 second, and the trial was aborted and omitted from analysis. A blue or pink rectangle appeared immediately afterward, indicating the final state. The participant then pressed the up-arrow key and, if the final state was rewarded, a green dollar sign appeared on the screen for 2 seconds; otherwise, a black X appeared for 2 seconds. The task comprised 250 trials, with a break every 50 trials, and participants received the total reward they obtained by the end of the task (0.18 CHF per reward).

### 4.3 Model-free algorithm

The SARSA model-free algorithm with replacing eligibility traces [1, 17] was used to simulate model-free learning agents. For each action  $a$  and state  $s$ , it estimated the value  $Q(s, a)$  of performing that action in that state. The task's initial states  $s_i$  were AA, AB, BA, and BB, and the actions  $a_i$  available at the initial states were *left* and *right*. The final states were *pink* and *blue*, and the only action  $a_f$  available at those states was *up*. The initial value of  $Q(s, a)$  for every state and action was 0.5. In each trial  $t$ , the simulated agent at the initial state  $s_i$  chose *left* as its initial-state action with probability  $p_{left}$  and *right* with probability  $1 - p_{left}$ , according to the following equation:

$$p_{left} = \frac{1}{1 + e^{-\beta[Q(s_i, left) - Q(s_i, right)]}}, \quad (1)$$

where  $\beta > 0$  is an inverse temperature parameter that determines the algorithm’s propensity to choose the option with the highest estimated value. After the final state  $s_f$  was observed and a reward  $r \in \{0, 1\}$  was received, state-action values were updated according to the following equations:

$$Q(s_i, a_i) = (1 - \alpha_1)Q(s_i, a_i) + \alpha_1 Q(s_f, up) + \alpha_1 \lambda [r - Q(s_f, up)], \quad (2)$$

$$Q(s_f, up) = (1 - \alpha_2)Q(s_f, up) + \alpha_2 r, \quad (3)$$

where  $0 \leq \alpha_1, \alpha_2, \lambda \leq 1$  are parameters:  $\alpha_1$  is the initial learning rate,  $\alpha_2$  is the final learning rate, and  $\lambda$  is the eligibility trace [1, 17].

In the special case where  $\lambda = 1$ , the update of initial state-action values becomes

$$Q(s_i, a_i) = (1 - \alpha_1)Q(s_i, a_i) + \alpha_1 r, \quad (4)$$

that is, the estimated values of choosing *left* and *right* in each initial state are updated independently of the final state’s estimated value. Thus, SARSA ( $\lambda = 1$ ) ignores the identity of the final state when making initial-state decisions, and an initial-state action that resulted in a reward will necessarily lead to a higher stay probability when the respective initial state recurs. This is true even if the action will probably lead to the final state with the lowest value.

#### 4.4 Model-based algorithm

In simulations of model-based agents [17], values were assigned to initial-state actions and to final states. The value  $V$  of a final state  $s \in \{pink, blue\}$  in the first trial  $t = 1$  was  $V(s, 1) = 0.5$ . An initial-state choice  $c \in \{left, right\}$  in trial  $t$  had a value  $V$  given by

$$V(c, t) = \Pr(c \rightarrow pink)V(pink, t) + \Pr(c \rightarrow blue)V(blue, t), \quad (5)$$

where  $\Pr(c \rightarrow s)$  is the probability that choosing  $c$  will lead to the final state  $s$ , which might be 0.8 or 0.2 according to the task’s transition model. The value of an initial-state choice can thus be understood as the expected value of the final state the agent will go to after making that choice. If  $V(left, t) > V(right, t)$ , the agent was more likely to choose left and vice-versa.

In each trial  $t$ , the agent’s initial state action was *left* with probability  $p_{left}$  and *right* with

probability  $1 - p_{left}$ , given by

$$p_{left} = \frac{1}{1 + e^{-\beta[V(left,t) - V(right,t)]}}, \quad (6)$$

where  $\beta$  is an inverse temperature parameter. After the agent made its initial-state choice and went to a final state  $s$ , that final state's value was updated according to the following equation:

$$V(s, t + 1) = (1 - \alpha)V(s, t) + \alpha r(t), \quad (7)$$

where  $r(t) \in \{0, 1\}$  indicates if the agent received a reward and  $0 \leq \alpha \leq 1$  is a learning-rate parameter of the model. The value of a final state is thus the moving average of the rewards received in that state.

#### 4.4.1 Model-based predictions

Our method of determining model-based predictions for the stay probability was different from the method used by Daw et al. [17]. In that study, there was only one initial state and the model-based and model-free algorithms predicted how the stay probability would change from one trial to the next. The present study's task, on the other hand, had four initial states and the model-free algorithm made predictions about how rewards would affect the participant's choices from one trial to the next trial *with the same initial state*, which is not necessarily the next trial. We therefore had to devise an alternative method of calculating the model-based predictions.

Our method relies directly on how the model-based algorithm estimates the reward probabilities of the initial-state choices, which either increase or decrease from one trial to the next with the same initial state depending on what happened, and was therefore learned about reward probabilities, in the intervening trials. If the participant's initial-state choice in a trial  $t_1$  was *left*, for instance, the model-based prediction was that in a future trial  $t_2$  with the same initial state the stay probability should increase if  $V(left, t_2) - V(right, t_2) > V(left, t_1) - V(right, t_1)$  and decrease otherwise. The model-based predictions depended on the parameter  $\alpha$ . The data analysis results were obtained by setting  $\alpha = 0.4$ , as this was the mean value that Daw et al. [17] found in their experiment by fitting to their experimental data an expanded reinforcement learning model that combines model-based and model-free learning. For comparison, we tried other values for  $\alpha$ , but the analysis results did not vary significantly.



## 399 4.5 Data analysis by logistic regression

400 For each human participant or simulated agent, we calculated the stay probability as a function of  
 401 model-free and model-based predictions. In each trial, if the human participant or simulated agent  
 402 chose an action that was the same as that chosen in the previous trial with the same initial state,  
 403 this was considered a stay. The four initial-state choices following the first occurrence of an initial  
 404 state were not analyzed. The remaining initial-state choices were coded as the random variable  $y$   
 405 and classified as a stay ( $y = 1$ ) or not a stay ( $y = 0$ ).

406 We then analyzed the resulting data using a hierarchical logistic regression model whose pa-  
 407 rameters were estimated through Bayesian computational methods. The dependent variable was  
 408  $p_{\text{stay}}$ , the stay probability for a given trial, and the independent variables were  $x_{mf}$ , which indi-  
 409 cated what the model-free algorithm predicted about  $p_{\text{stay}}$  (+1 if it predicted an increase, -1 if it  
 410 predicted a decrease),  $x_{mb}$ , which indicated what the model-based algorithm predicted about  $p_{\text{stay}}$   
 411 (+1 if it predicted an increase, -1 if it predicted a decrease), and the interaction between the two.  
 412 Thus, for each participant, we determined a four-dimensional vector  $\vec{\beta}$  whose components were the  
 413  $\beta$  coefficients of the following equation:

$$p_{\text{stay}} = \frac{1}{1 + \exp[-(\beta_0 + \beta_{mb}x_{mb} + \beta_{mf}x_{mf} + \beta_{mb \times mf}x_{mf}x_{mb})]}. \quad (8)$$

414 The distribution of  $y$  was Bernoulli( $p_{\text{stay}}$ ). The distribution of the  $\vec{\beta}$  vectors was  $\mathcal{N}(\vec{\mu}_c, \vec{\sigma}^2)$  if  
 415 the participant was in the simultaneous condition and  $\mathcal{N}(\vec{\mu}_e, \vec{\sigma}^2)$  if the participant was in the  
 416 sequential condition; in other words, the group means for each  $\vec{\beta}$  were allowed to vary independently.  
 417 The parameters of the  $\vec{\beta}$  distribution were given vague prior distributions based on preliminary  
 418 analyses—the  $\vec{\mu}$  vectors' components were given a  $\mathcal{N}(\mu = 0, \sigma^2 = 25)$  prior, and the  $\vec{\sigma}^2$  vector's  
 419 components were given a Half-normal(0, 25) prior. Other vague prior distributions for the model  
 420 parameters were tested and the results did not change significantly.

421 To obtain parameter estimates from the model's posterior distribution, we coded the model  
 422 into the Stan modeling language version 2.14.0 [39, 40] and used the PyStan Python package [41]  
 423 to obtain 100,000 samples of the joint posterior distribution from four chains of length 50,000  
 424 (warmup 25,000). Convergence of the chains was indicated by  $\hat{R} \approx 1.0$  for all parameters. The  
 425 minimum effective sample size for the parameters of interest  $\vec{\mu}_c$ ,  $\vec{\mu}_e$ , and  $\vec{\mu}_e - \vec{\mu}_c$  was 31785.

## 426 4.6 Fitting of the algorithms to experimental data

427 For comparison with the participant data, we fitted the SARSA model-free algorithm and the  
 428 model-based algorithm to the experimental data and generated replicated data using the fitted pa-

rameters. The parameters were obtained by fitting both algorithms to all participants (to generate Figures 3 and 4) and the model-based algorithm to the participants in the sequential condition (to perform the simulated experiments). To that end, we used a Bayesian hierarchical model, which allowed us to pool data from all participants to improve individual parameter estimates.

The parameters of the model-based algorithm for the  $i$ th participant were  $\alpha^i$  and  $\beta^i$ . They were given a  $\text{Beta}(a_\alpha, b_\alpha)$  and  $\ln \mathcal{N}(\mu_\beta, \sigma_\beta^2)$  prior distributions respectively. The hyperparameters  $a_\alpha$  and  $b_\alpha$  were themselves given a noninformative  $\text{Half-normal}(0, 10^4)$  prior and the hyperparameters  $\mu_\beta$  and  $\sigma_\beta^2$  were given a noninformative  $\mathcal{N}(0, 10^4)$  and  $\text{Half-normal}(0, 10^4)$  priors respectively. The parameters of the model-free algorithm for the  $i$ th participant were  $\alpha_1^i$ ,  $\alpha_2^i$ ,  $\lambda^i$ , and  $\beta^i$ . They were given a  $\text{Beta}(a_{\alpha_1}, b_{\alpha_1})$ ,  $\text{Beta}(a_{\alpha_2}, b_{\alpha_2})$ ,  $\text{Beta}(a_\lambda, b_\lambda)$  and  $\ln \mathcal{N}(\mu_\beta, \sigma_\beta^2)$  prior distributions respectively. The hyperparameters  $a_{\alpha_1}$ ,  $a_{\alpha_2}$ ,  $a_\lambda$ ,  $b_{\alpha_1}$ ,  $b_{\alpha_2}$ , and  $b_\lambda$  were themselves given a noninformative  $\text{Half-normal}(0, 10^4)$  prior and the hyperparameters  $\mu_\beta$  and  $\sigma_\beta^2$  were given a noninformative  $\mathcal{N}(0, 10^4)$  and  $\text{Half-normal}(0, 10^4)$  priors respectively. We then coded the models into the Stan modeling language version 2.14.0 [39, 40] and used the PyStan Python package [41] to obtain 50,000 samples of the joint posterior distribution from one chain of length 60,000 (warmup 10,000). Convergence of the chains was indicated by  $\hat{R} \approx 1.0$  for all parameters. The minimum effective sample size was 1481 for all hyperparameters. The results were used to generate Figures 2 and 3.

## 4.7 Simulated experiments

Given that this study's aim was to determine if working memory-dependent temporal pattern learning is necessarily model-based or can be model-free, we sought to determine if the results obtained for the sequential condition could have been generated by the model-based algorithm. To this end, we simulated 10,000 experiments wherein, in each simulated experiment, the 21 participants in the simultaneous condition were compared to a different group of 20 simulated purely-model-based agents (as replacements for the 20 human participants in the sequential condition).

The model-based algorithm was first fitted to the sequential condition results using the Bayesian hierarchical method described above to obtain 200,000 samples of the posterior distribution from four chains of length 60,000 (warmup 10,000). Convergence of the chains was indicated by  $\hat{R} \approx 1.0$  for all parameters. The minimum effective sample size was 16467 for all hyperparameters. For each simulated experiment, a point was randomly selected from the posterior distribution of hyperparameters  $(a_\alpha, b_\alpha, \mu_\beta, \sigma_\beta^2)$  and 20 sets of algorithm parameters  $(\alpha, \beta)$  were randomly generated using the selected values, i.e.  $\alpha \sim \text{Beta}(a_\alpha, b_\alpha)$ ,  $\beta \sim \ln \mathcal{N}(\mu_\beta, \sigma_\beta^2)$ . For each  $(\alpha, \beta)$  parameter set, the model-based algorithm was run for 250 trials of the experimental task to generate

462 results for a simulated purely-model-based agent. These simulated agents were then compared with  
463 the actual participants in the simultaneous condition using the same logistic regression analysis  
464 described above, except that, for computational efficiency, only 600 samples from one chain of 800  
465 samples (warmup 200) was obtained from the posterior distribution.

466 The entire analysis procedure was replicated several times with differing parameter values and  
467 prior distributions to ensure that the results and conclusions remained the same under a wide set  
468 of assumptions. In all cases, the results were nearly identical and supported the same conclusions.

## 469 4.8 Code and data availability

470 All the computer code and behavioral data used in this study are available at [https://github.com/carolfs/mf\\_wm](https://github.com/carolfs/mf_wm)

## 471 5 Acknowledgements

472 This work was supported by the São Paulo Research Foundation – FAPESP (grant number  
473 2013/10694-0) and the start-up research funds from the University of Zurich. Y.Y.’s involvement  
474 was supported by the China Scholarship Council.

## 475 6 Author contributions

476 C.F.S. and T.A.H. designed the study; C.F.S. and Y.Y. conducted the behavioral experiment;  
477 C.F.S. performed the simulations and analyzed the data with input from T.A.H.; C.F.S., Y.Y.,  
478 and T.A.H. wrote the manuscript.

## 479 7 Competing financial interests

480 The authors declare no competing financial interests.

## 481 References

- 482 [1] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. A  
483 Bradford Book, first edition, 1998.
- 484 [2] W. Schultz, P. Dayan, and P. R. Montague. A Neural Substrate of Prediction and Reward.  
485 *Science*, 275(5306):1593–1599, mar 1997. ISSN 0036-8075. doi: 10.1126/science.275.5306.1593.  
486 URL <http://www.sciencemag.org/cgi/doi/10.1126/science.275.5306.1593>.

- 487 [3] Christopher D Fiorillo, Philippe N Tobler, and Wolfram Schultz. Discrete coding of reward  
488 probability and uncertainty by dopamine neurons. *Science (New York, N.Y.)*, 299(5614):1898–  
489 902, mar 2003. ISSN 1095-9203. doi: 10.1126/science.1077349. URL [http://www.ncbi.nlm.  
490 nih.gov/pubmed/12649484](http://www.ncbi.nlm.nih.gov/pubmed/12649484).
- 491 [4] Yael Niv. Reinforcement learning in the brain. *Journal of Mathematical Psychology*, 53(3):139–  
492 154, jun 2009. ISSN 00222496. doi: 10.1016/j.jmp.2008.12.005. URL [http://linkinghub.  
493 elsevier.com/retrieve/pii/S0022249608001181](http://linkinghub.elsevier.com/retrieve/pii/S0022249608001181).
- 494 [5] P. W. Glimcher. Understanding dopamine and reinforcement learning: The dopamine re-  
495 ward prediction error hypothesis. *Proceedings of the National Academy of Sciences*, 108  
496 (Supplement\_3):15647–15654, sep 2011. ISSN 0027-8424. doi: 10.1073/pnas.1014269108.  
497 URL <http://www.pnas.org/cgi/doi/10.1073/pnas.1014269108>.
- 498 [6] Daeyeol Lee, Hyojung Seo, and Min Whan Jung. Neural Basis of Reinforcement Learning  
499 and Decision Making. *Annual Review of Neuroscience*, 35(1):287–308, jul 2012. ISSN 0147-  
500 006X. doi: 10.1146/annurev-neuro-062111-150512. URL [http://www.annualreviews.org/  
501 doi/abs/10.1146/annurev-neuro-062111-150512](http://www.annualreviews.org/doi/abs/10.1146/annurev-neuro-062111-150512).
- 502 [7] Ray J. Dolan and Peter Dayan. Goals and Habits in the Brain. *Neuron*, 80(2):312–325,  
503 oct 2013. ISSN 08966273. doi: 10.1016/j.neuron.2013.09.007. URL [http://linkinghub.  
504 elsevier.com/retrieve/pii/S0896627313008052](http://linkinghub.elsevier.com/retrieve/pii/S0896627313008052).
- 505 [8] Ann M. Graybiel. Habits, Rituals, and the Evaluative Brain. *Annual Review of Neuroscience*,  
506 31(1):359–387, jul 2008. ISSN 0147-006X. doi: 10.1146/annurev.neuro.29.051605.112851. URL  
507 <http://www.annualreviews.org/doi/10.1146/annurev.neuro.29.051605.112851>.
- 508 [9] Peter Dayan. How to set the switches on this thing. *Current Opinion in Neurobiology*,  
509 22(6):1068–1074, dec 2012. ISSN 09594388. doi: 10.1016/j.conb.2012.05.011. URL [http:  
510 //linkinghub.elsevier.com/retrieve/pii/S0959438812000992](http://linkinghub.elsevier.com/retrieve/pii/S0959438812000992).
- 511 [10] Kyle S. Smith and Ann M. Graybiel. Investigating habits: strategies, technologies and mod-  
512 els. *Frontiers in Behavioral Neuroscience*, 8, 2014. ISSN 1662-5153. doi: 10.3389/fnbeh.2014.  
513 00039. URL [http://journal.frontiersin.org/article/10.3389/fnbeh.2014.00039/  
514 abstract](http://journal.frontiersin.org/article/10.3389/fnbeh.2014.00039/abstract).
- 515 [11] Fiery Cushman and Adam Morris. Habitual control of goal selection in humans. *Pro-  
516 ceedings of the National Academy of Sciences*, 112(45):13817–13822, nov 2015. ISSN 0027-

8424. doi: 10.1073/pnas.1506367112. URL <http://www.pnas.org/lookup/doi/10.1073/pnas.1506367112>.

[12] Henk Aarts and Ap Dijksterhuis. Habits as knowledge structures: Automaticity in goal-directed behavior. *Journal of Personality and Social Psychology*, 78(1):53–63, 2000. ISSN 1939-1315. doi: 10.1037/0022-3514.78.1.53. URL <http://doi.apa.org/getdoi.cfm?doi=10.1037/0022-3514.78.1.53>.

[13] Amir Dezfouli and Bernard W. Balleine. Habits, action sequences and reinforcement learning. *European Journal of Neuroscience*, 35(7):1036–1051, apr 2012. ISSN 0953816X. doi: 10.1111/j.1460-9568.2012.08050.x. URL <http://doi.wiley.com/10.1111/j.1460-9568.2012.08050.x>.

[14] Amir Dezfouli and Bernard W. Balleine. Actions, Action Sequences and Habits: Evidence That Goal-Directed and Habitual Action Control Are Hierarchically Organized. *PLoS Computational Biology*, 9(12):e1003364, dec 2013. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003364. URL <http://dx.plos.org/10.1371/journal.pcbi.1003364>.

[15] A. Dezfouli, N. W. Lingawi, and B. W. Balleine. Habits as action sequences: hierarchical action control and changes in outcome value. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1655):20130482–20130482, sep 2014. ISSN 0962-8436. doi: 10.1098/rstb.2013.0482. URL <http://rstb.royalsocietypublishing.org/cgi/doi/10.1098/rstb.2013.0482>.

[16] Randall C. O’Reilly and Michael J. Frank. Making Working Memory Work: A Computational Model of Learning in the Prefrontal Cortex and Basal Ganglia. *Neural Computation*, 18(2):283–328, feb 2006. ISSN 0899-7667. doi: 10.1162/089976606775093909. URL <http://www.mitpressjournals.org/doi/abs/10.1162/089976606775093909>.

[17] Nathaniel D. Daw, Samuel J. Gershman, Ben Seymour, Peter Dayan, and Raymond J. Dolan. Model-Based Influences on Humans’ Choices and Striatal Prediction Errors. *Neuron*, 69(6):1204–1215, mar 2011. ISSN 08966273. doi: 10.1016/j.neuron.2011.02.027. URL [http://www.cell.com/neuron/abstract/S0896-6273\(11\)00125-5](http://www.cell.com/neuron/abstract/S0896-6273(11)00125-5)<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3077926&tool=pmcentrez&rendertype=abstract><http://linkinghub.elsevier.com/retrieve/pii/S0896627311001255>.

[18] Michael T Todd, Yael Niv, and Jonathan D Cohen. Learning to Use Working Memory in Partially Observable Environments through Dopaminergic Reinforcement. In D Koller, D Schuurmans, Y Bengio, and L Bottou, editors, *Advances in Neural Information Processing Systems*

21, pages 1689–1696. Curran Associates, Inc., 2009. URL <http://papers.nips.cc/paper/3508-learning-to-use-working-memory-in-partially-observable-environments-through-dopaminergic-learning>.pdf.

[19] Arthur S. Reber. Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General*, 118(3):219–235, 1989. ISSN 1939-2222. doi: 10.1037/0096-3445.118.3.219. URL <http://doi.apa.org/getdoi.cfm?doi=10.1037/0096-3445.118.3.219>.

[20] Richard L. Canfield and Marshall M. Haith. Young infants’ visual expectations for symmetric and asymmetric stimulus sequences. *Developmental Psychology*, 27(2):198–208, 1991. ISSN 0012-1649. doi: 10.1037/0012-1649.27.2.198. URL <http://cat.inist.fr/?aModele=afficheN&cpsidt=19452330><http://doi.apa.org/getdoi.cfm?doi=10.1037/0012-1649.27.2.198>.

[21] Scott A. Huettel, Peter B. Mack, and Gregory McCarthy. Perceiving patterns in random series: dynamic processing of sequence in prefrontal cortex. *Nature Neuroscience*, apr 2002. ISSN 10976256. doi: 10.1038/nn841. URL <http://www.nature.com/doifinder/10.1038/nn841>.

[22] Asher Cohen, Richard I. Ivry, and Steven W. Keele. Attention and structure in sequence learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(1):17–30, 1990. ISSN 1939-1285. doi: 10.1037/0278-7393.16.1.17. URL <http://doi.apa.org/getdoi.cfm?doi=10.1037/0278-7393.16.1.17>.

[23] Axel Cleeremans and James L. McClelland. Learning the structure of event sequences. *Journal of Experimental Psychology: General*, 120(3):235–253, 1991. ISSN 1939-2222. doi: 10.1037/0096-3445.120.3.235. URL <http://doi.apa.org/getdoi.cfm?doi=10.1037/0096-3445.120.3.235>.

[24] I H Jenkins, D J Brooks, P D Nixon, R S Frackowiak, and R E Passingham. Motor sequence learning: a study with positron emission tomography. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 14(6):3775–90, jun 1994. ISSN 0270-6474. URL <http://www.ncbi.nlm.nih.gov/pubmed/8207487>.

[25] Eli Vakil, Shimon Kahan, Moshe Huberman, and Alicia Osimani. Motor and non-motor sequence learning in patients with basal ganglia lesions: The case of serial reaction time (SRT). *Neuropsychologia*, 38(1):1–10, 2000. ISSN 00283932. doi: 10.1016/S0028-3932(99)00058-5.

[26] S. Lehericy, H. Benali, P.-F. Van de Moortele, M. Pelegrini-Issac, T. Waechter, K. Ugurbil, and J. Doyon. Distinct basal ganglia territories are engaged in early and advanced motor

- sequence learning. *Proceedings of the National Academy of Sciences*, 102(35):12566–12571,  
aug 2005. ISSN 0027-8424. doi: 10.1073/pnas.0502762102. URL [http://www.pnas.org/  
cgi/doi/10.1073/pnas.0502762102](http://www.pnas.org/cgi/doi/10.1073/pnas.0502762102).
- [27] A. R. Otto, C. M. Raio, A. Chiang, E. A. Phelps, and N. D. Daw. Working-memory capacity  
protects model-based learning from stress. *Proceedings of the National Academy of Sciences*,  
110(52):20941–20946, dec 2013. ISSN 0027-8424. doi: 10.1073/pnas.1312011110. URL [http:  
//www.pnas.org/cgi/doi/10.1073/pnas.1312011110](http://www.pnas.org/cgi/doi/10.1073/pnas.1312011110).
- [28] A. Ross Otto, Samuel J. Gershman, Arthur B. Markman, and Nathaniel D. Daw. The Curse  
of Planning. *Psychological Science*, 24(5):751–761, may 2013. ISSN 0956-7976. doi: 10.1177/  
0956797612463080. URL <http://journals.sagepub.com/doi/10.1177/0956797612463080>.
- [29] A. Ross Otto, Anya Skatova, Seth Madlon-Kay, and Nathaniel D. Daw. Cognitive Control  
Predicts Use of Model-based Reinforcement Learning. *Journal of Cognitive Neuroscience*, 27  
(2):319–333, feb 2015. ISSN 0898-929X. doi: 10.1162/jocn\_a\_00709. URL [http://www.  
mitpressjournals.org/doi/abs/10.1162/jocn\\_a\\_00709](http://www.mitpressjournals.org/doi/abs/10.1162/jocn_a_00709).
- [30] J. H. Decker, A. R. Otto, N. D. Daw, and C. A. Hartley. From Creatures of Habit  
to Goal-Directed Learners: Tracking the Developmental Emergence of Model-Based Re-  
inforcement Learning. *Psychological Science*, 27(6):848–858, jun 2016. ISSN 0956-7976.  
doi: 10.1177/0956797616639301. URL [http://pss.sagepub.com/lookup/doi/10.1177/  
0956797616639301](http://pss.sagepub.com/lookup/doi/10.1177/0956797616639301).
- [31] Eva Friedel, Stefan P. Koch, Jean Wendt, Andreas Heinz, Lorenz Deserno, and Flo-  
rian Schlagenhauf. Devaluation and sequential decisions: linking goal-directed and model-  
based behavior. *Frontiers in Human Neuroscience*, 8, aug 2014. ISSN 1662-5161. doi:  
10.3389/fnhum.2014.00587. URL [http://journal.frontiersin.org/article/10.3389/  
fnhum.2014.00587/abstract](http://journal.frontiersin.org/article/10.3389/fnhum.2014.00587/abstract).
- [32] Samuel J. Gershman, Arthur B. Markman, and A. Ross Otto. Retrospective revaluation  
in sequential decision making: A tale of two systems. *Journal of Experimental Psychology:  
General*, 143(1):182–194, 2014. ISSN 1939-2222. doi: 10.1037/a0030844. URL [http://doi.  
apa.org/getdoi.cfm?doi=10.1037/a0030844](http://doi.apa.org/getdoi.cfm?doi=10.1037/a0030844).
- [33] John P. O’Doherty, Jeffrey Cockburn, and Wolfgang M. Pauli. Learning, Reward, and Decision  
Making. *Annual Review of Psychology*, 68(1):73–100, jan 2017. ISSN 0066-4308. doi: 10.  
1146/annurev-psych-010416-044216. URL [http://www.annualreviews.org/doi/10.1146/  
annurev-psych-010416-044216](http://www.annualreviews.org/doi/10.1146/annurev-psych-010416-044216).



- [34] P. Read Montague, Raymond J. Dolan, Karl J. Friston, and Peter Dayan. Computational psychiatry. *Trends in Cognitive Sciences*, 16(1):72–80, jan 2012. ISSN 13646613. doi: 10.1016/j.tics.2011.11.018. URL <http://linkinghub.elsevier.com/retrieve/pii/S1364661311002518>.
- [35] G. Cumming. Precision for Planning. In *Understanding The New Statistics*, chapter 13, pages 355–380. Routledge, New York, London, 1 edition, 2012.
- [36] J. K. Kruschke. Goals, Power, and Sample Size. In *Doing Bayesian Data Analysis*, chapter 13, pages 359–398. Academic Press, London, 2 edition, 2015.
- [37] Nathaniel D Daw, Yael Niv, and Peter Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience*, 8(12):1704–11, dec 2005. ISSN 1097-6256. doi: 10.1038/nn1560. URL <http://dx.doi.org/10.1038/nn1560><http://www.ncbi.nlm.nih.gov/pubmed/16286932>.
- [38] Nathaniel D. Daw and John P. O’Doherty. Multiple Systems for Value Learning. In Paul W. Glimcher and Ernst Fehr, editors, *Neuroeconomics*, chapter 21, pages 393–410. Elsevier, second edition, 2014. doi: 10.1016/B978-0-12-416008-8.00021-8. URL <http://linkinghub.elsevier.com/retrieve/pii/B9780124160088000218>.
- [39] Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan : A Probabilistic Programming Language. *Journal of Statistical Software*, 76(1), 2017. ISSN 1548-7660. doi: 10.18637/jss.v076.i01. URL <http://www.jstatsoft.org/v76/i01/>.
- [40] Stan Development Team. Stan Modeling Language Users Guide and Reference Manual, Version 2.14.0, 2016.
- [41] Stan Development Team. PyStan: the Python interface to Stan, 2016. URL <http://mc-stan.org>.